NYPD_Shooting_Project

Ryan Scott

2025-06-20

NYPD Shooting Incident Project

Statement of Interest

The following project is an Exploratory Data Analysis of NYPD Shooting Incident Data from 2006 to 2024. The questions we want to answer are:

What year had the highest number of shooting incidents? Which Boroughs had the highest number of shooting incidents?

Background

We were able to access the files for this project at https://catalog.data.gov/dataset and find the dataset titled NYPD Shooting Incident Data (Historic).

To begin, I install the appropriate R packages.

```
#install.packages("tidyverse")
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Warning: package 'purrr' was built under R version 4.3.3
```

```
## Warning: package 'lubridate' was built under R version 4.3.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

I then assign the NYPD csv file to a variable in Rstudio.

```
NYPD <- read_csv("~/Desktop/Documents/Data Analysis Tools/Projects/RStudio Projects/NYPD/NYPD_Shooting_
```

```
## Rows: 29744 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (5): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, Latitude, Longitude
## num   (2): X_COORD_CD, Y_COORD_CD
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

I want to get an idea of the columns I will be working with, so I use the head function and set the number of rows to 5.

```
head(NYPD, n = 5)
```

```
## # A tibble: 5 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>    <chr>                <dbl>
## 1    231974218 08/09/2021 01:06      BRONX    <NA>                    40
## 2    177934247 04/07/2018 19:48      BROOKLYN <NA>                    79
## 3    255028563 12/02/2022 22:57      BRONX    OUTSIDE                 47
## 4     25384540 11/19/2006 01:50      BROOKLYN <NA>                    66
## 5     72616285 05/09/2010 01:58      BRONX    <NA>                    46
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <chr>
```

I use the summary function to identify the types of data in each column.

```
summary(NYPD)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:29744       Length:29744       Length:29744
##  1st Qu.: 67321140   Class :character   Class1:hms         Class :character
##  Median :109291972   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :133850951                      Mode  :numeric
##  3rd Qu.:214741917
##  Max.   :299462478
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:29744       Min.   :  1.00   Min.   :0.0000    Length:29744
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.00   Median :0.0000    Mode  :character
##                     Mean   : 65.23   Mean   :0.3181
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
```

```
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:29744       Mode :logical           Length:29744
##  Class :character   FALSE:23979             Class :character
##  Mode  :character   TRUE :5765              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE          VIC_AGE_GROUP        VIC_SEX
##  Length:29744       Length:29744       Length:29744       Length:29744
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE             X_COORD_CD        Y_COORD_CD        Latitude
##  Length:29744       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000094   1st Qu.:183042   1st Qu.:40.67
##  Mode  :character   Median :1007826   Median :195506   Median :40.70
##                     Mean   :1009442   Mean   :208722   Mean   :40.74
##                     3rd Qu.:1016739   3rd Qu.:239980   3rd Qu.:40.83
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :97
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:29744
##  1st Qu.:-73.94   Class :character
##  Median :-73.91   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :97
```

After reviewing the columns and the data types of the columns, I change Occurence Date to date format from character and remore the Latitude, Longitude, and Coordinate columns from the table

```
NYPD <- NYPD %>%
    mutate(OCCUR_Date = as.Date(OCCUR_DATE, format = "%m/%d/%Y"))
```

```
NYPD <- NYPD %>%
     select(-Latitude, -Longitude, -Lon_Lat, -Y_COORD_CD, -X_COORD_CD)
```

I want to see the date range of the data, so I check the minimun and maximum dates of the Occurence Date column.

```
date_summary <- NYPD %>%
     summarise(
         max_occur_date = max(OCCUR_Date, na.rm = TRUE),
         min_occur_date = min(OCCUR_Date, na.rm = TRUE)
     )

print(date_summary)
```

```
## # A tibble: 1 x 2
##   max_occur_date min_occur_date
##   <date>         <date>
## 1 2024-12-31     2006-01-01
```

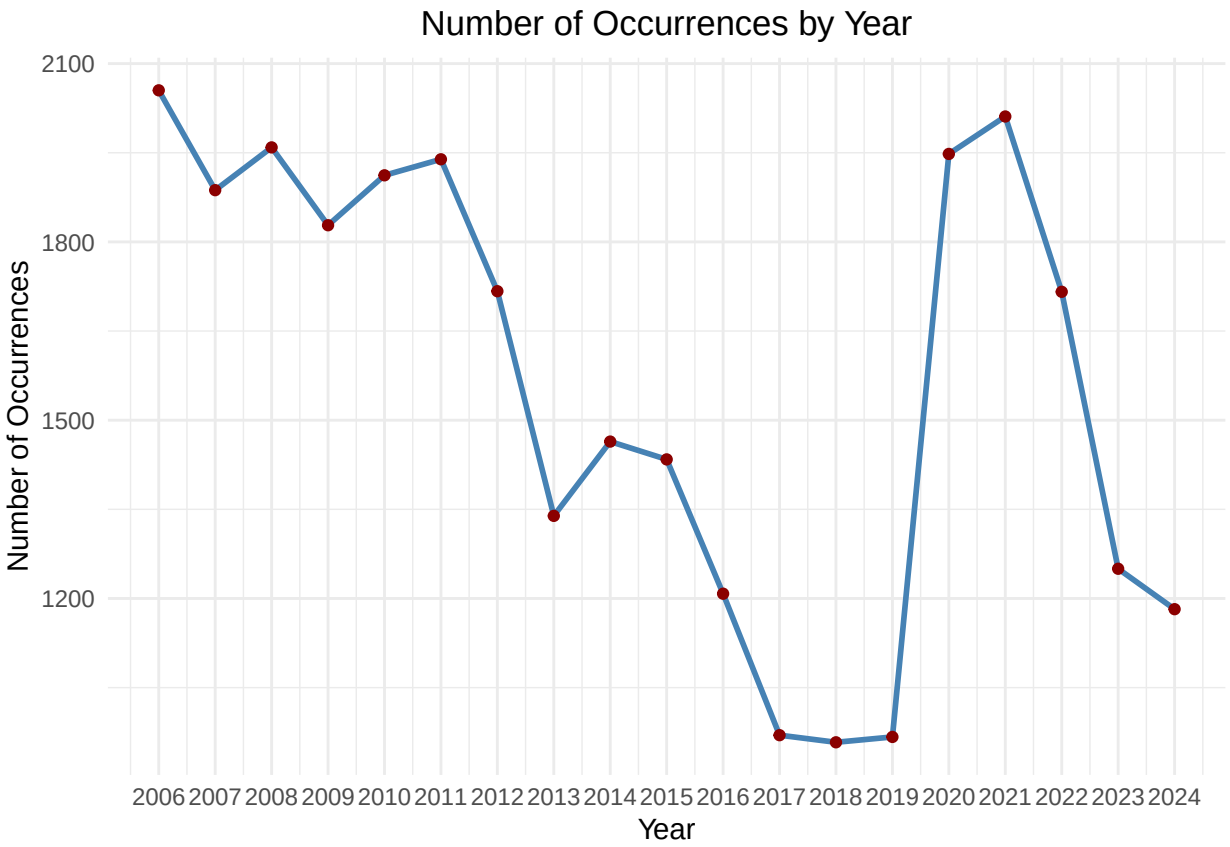
```r
occurrences_by_year <- NYPD %>%
  mutate(Occur_Year = format(OCCUR_Date, "%Y")) %>%
  group_by(Occur_Year) %>%
  summarise(Count = n()) %>%
  ungroup() %>%
  mutate(Occur_Year = as.numeric(Occur_Year)) %>%
  arrange(Occur_Year)
```

From this code, I was able to determine that the date range is from 01-01-2006 to 12-31-2024.

The following visualization shows the number of occurrences by Year in ascending order.

```r
ggplot(occurrences_by_year, aes(x = Occur_Year, y = Count)) +
     geom_line(color = "steelblue", linewidth = 1) +
     geom_point(color = "darkred", linewidth = 2) +
     labs(
         title = "Number of Occurrences by Year",
         x = "Year",
         y = "Number of Occurrences"
     ) +
     theme_minimal() +
     theme(plot.title = element_text(hjust = 0.5)) +
     scale_x_continuous(breaks = unique(occurrences_by_year$Occur_Year))
```

```
## Warning in geom_point(color = "darkred", linewidth = 2): Ignoring unknown
## parameters: ‘linewidth‘
```

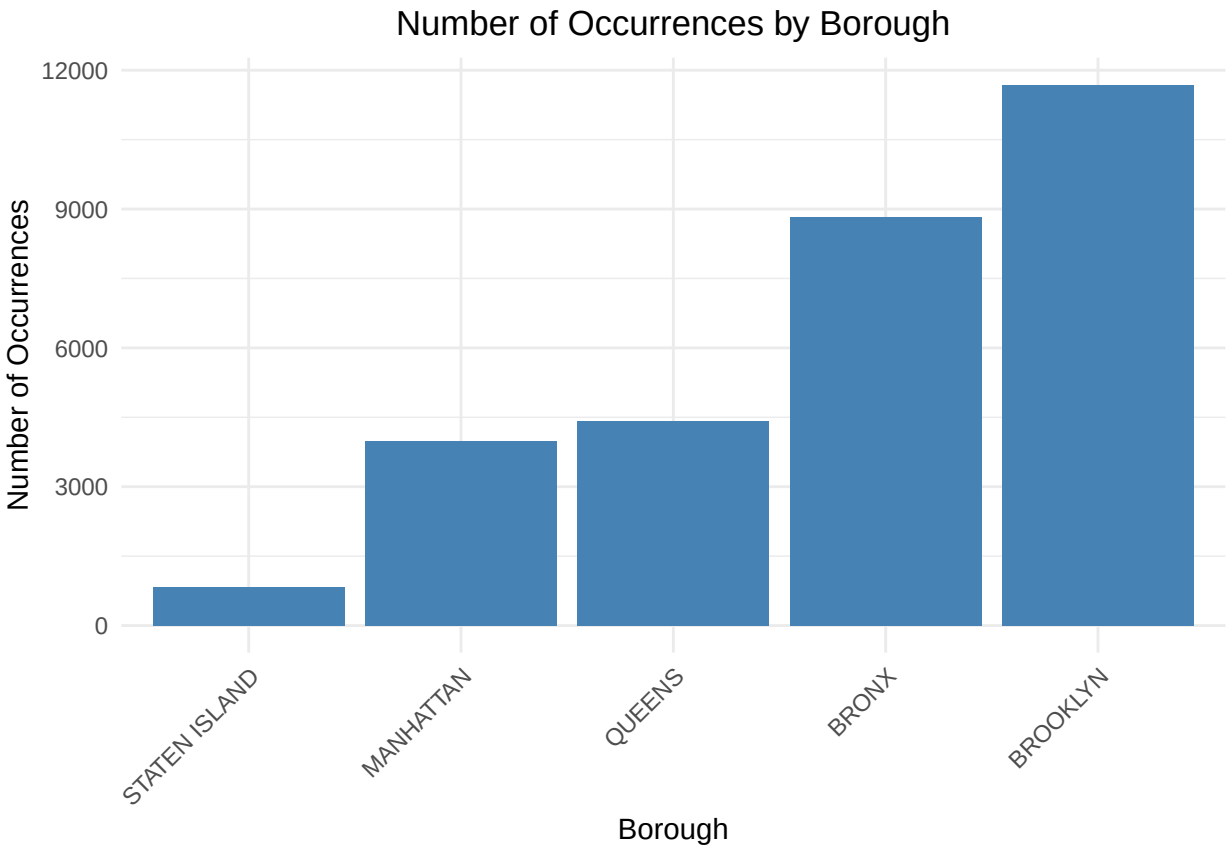


From the previous visualization, I am able to determine that with this date range, 2006 had the highest number of occurrences.I am also able to determine that number of occurrences fell from 2006 to 2019 before rising sharply in 2020.

```
borough_counts <- NYPD %>%
  count(BORO) %>%
  mutate(BORO = fct_reorder(BORO, n))
```

The following visualization shows the number of occurrences by New York City Borough.

```
ggplot(borough_counts, aes(x = BORO, y = n)) +
     geom_col(fill = "steelblue") +
     labs(
         title = "Number of Occurrences by Borough",
         x = "Borough",
         y = "Number of Occurrences"
     ) +
     theme_minimal() +
     theme(
         plot.title = element_text(hjust = 0.5), # Center the title
         axis.text.x = element_text(angle = 45, hjust = 1)
     )
```



After reviewing the previous visualization, I can determine that within the date range of 2006 to 2024, Brooklyn had the highest number of occurrences.

Conclusion

Upon conclusion, we were able to determine that 2006 had the highest number of NYPD Shooting Incidents and that Brooklyn had the highest number of shooting incidents.

Bias

I could not identify any source of bias within this project. I kept the questions simple and open ended to elimate any possible bias. As far as the data goes, the columns related to facts about the incidents, such as time, location, and date, so bias would not be present.